

# Visual Models of Interaction

David Hogg, Neil Johnson, Richard Morris, Dietrich Buesching, Aphrodite Galata  
School of Computer Studies

University of Leeds

Leeds, LS2 9JT, UK

{dch,neilj,rjm,bueschin,afro}@scs.leeds.ac.uk

<http://www.scs.leeds.ac.uk/vision>

## Abstract

An important research challenge is to find ways for acquiring and encoding the spatial, temporal and procedural variations between and within different types of human interaction. We describe recent work on interaction modelling in which the range of possible things that can happen is learnt automatically through passive observation of video sequences depicting typical interactions. The basis for the approach is the construction of probabilistic spatio-temporal models from training data extracted from the video sequences.

Results will be presented for two kinds of application of this work. The first is concerned with interactions between a pair of individuals with application to human-computer interaction. The second is dealing with interactions between people and motor vehicles, intended for wide-area surveillance.

## Introduction

For many years, researchers have sought to build intelligent agents able to interact with people in natural ways. With the advent of fast computer graphics, there is an understandable move to give these agents realistic bodies and faces on the computer screen. There is a similar requirement from the TV and film industry for the generation of realistic looking faces for the purpose of animation.

Several approaches to face generation have been developed. A 3-D texture rendered surface model of the head and face, combined with the ability to adjust the local shape of the surface (e.g. through moving individual vertices) provides expressive authenticity and the ability to move and rotate a face in 3-D with ease. Even greater realism has been obtained by simulating facial muscles and skin covering (Terzopoulos and Waters, 1990). An alternative approach has been to model the 2-D appearance of a face along with parameterised texture variations to simulate different individuals, and 2-D deformations to simulate different expressions and viewpoints (Edwards et al., 1998).

Although such models provide a very good graphical representation of facial appearance and facial motion in making expressions and in speaking, research on the generation of appropriate behaviour is less well understood. At a high level, talking heads have been made to produce speech utterances in response to spoken queries, but the mechanism driving facial movements is programmed off-line giving an unnatural appearance to the agent and a coarsely co-ordinated interaction.

In this paper, we review recent work that is addressing these problems by attempting to control a virtual person directly from a visual model of interaction learnt by observing real interactions. The potential benefit of this approach is that the sequence of body movements and timing in coordination with the user will appear much more realistic since they are derived directly from a compressed visual representation of the real thing.

The paper is in two parts. The first part describes the way in which interactive behaviours are learnt and then used for generating a virtual person with which to interact. The second part looks at the application of a related statistical learning procedure in making sense of the interactions between people and cars, with applications in surveillance.

## Models of Behaviour

We begin by describing a general method for representing the spatio-temporal evolution of a given target activity (from Johnson and Hogg, 1996). Such representations (referred to as *models*) are learnt automatically from long video segments (or *video corpora* by analogy with speech corpora in computational linguistics) depicting repeated examples of the target activity. Typically these video segments may contain many hundreds of examples that collectively span the range of ways in which the activity may be carried out.

To illustrate the method, we look at modelling the way in which pedestrians traverse a scene, in terms of their progressive location and speed on the ground-plane. The constructed models should be sufficiently detailed and accurate to serve as a reference from which the path of a pedestrian may be extrapolated forwards by several seconds or assessed for its typicality with respect to the pathways in the video corpus.

Such a system could have many uses. In a surveillance application, it could raise the alarm when someone moves atypically through a scene. It could also help to resolve long occlusions that occur when someone is lost from view due to a temporary obstacle in the scene. Short-range predictions from a dynamical model, such as a second order differential equation, do not provide the required accuracy that would for example be needed to cope with occlusions of a pathway in which there is a sharp turn.

### Obtaining training data

To train the behaviour model, a video segment of a pedestrian scene lasting several hours is recorded for subsequent processing off-line. The visual tracker developed by Baumberg and Hogg (1994, 1996) is used to extract the trajectories of each person traversing the viewed area. This tracker is based on the active shape models of Cootes, Taylor, Cooper, and Graham (1995) and is also closely related to the active contours of Blake and Isard (1998). Image profiles are modelled by a closed B-spline contour, as shown around both of the profiles in Figure 7. The contour is represented by concatenating the control points into a single vector (referred to as a *shape vector*):

$$(x_1, y_1, x_2, y_2, \dots, x_n, y_n)$$

For the profiles in Figure 7, this shape vector has 64 components ( $2 \times 32$ ). It is supposed a *training set* of typical profiles of a target shape class is available and that these aligned with one another. For the pedestrian tracker, this training set is obtained by a simple picture differencing strategy to segment moving objects from the background. A B-spline contour is wrapped around each extracted profile, with control points equally spaced, and arranged so that the first control point is located in a similar position on every profile (e.g. at the top). Thus, for each of  $m$  training profiles there is a shape vector of the form:

$$\mathbf{x}^i = (x_1^i, y_1^i, x_2^i, y_2^i, \dots, x_n^i, y_n^i) \quad i = 1, m$$

A re-parameterisation of this cluster of shape vectors is obtained using a principal component analysis about the mean profile, given by:

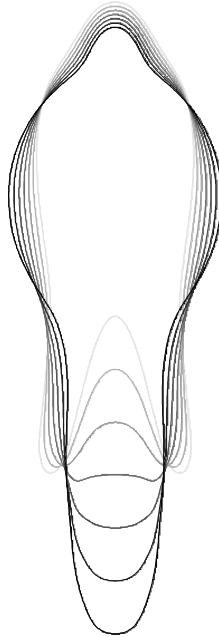
$$\bar{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^m \mathbf{x}^i$$

The aim of this analysis is to find a concise parameterisation of the cluster which in turn provides a model for the target shape with many fewer parameters than the dimensionality of the shape vectors. Assuming

most of the variation in the shape is accommodated by the  $t$  principal components with the largest eigenvalues, the parameterised model shape  $\mathbf{s}(\cdot)$  has the form:

$$\mathbf{s}(b_1, b_2, \dots, b_t) = \bar{\mathbf{x}} + b_1 \mathbf{v}_1 + b_2 \mathbf{v}_2 + \dots + b_t \mathbf{v}_t$$

For pedestrian tracking, as few as 10 principal components may be sufficient. Figure 1 shows the variation about the mean shape along the first principal component for the pedestrian model used in our tracker. The most noticeable feature of the variation is the characteristic appearance and disappearance of the gap between the legs.



*Figure 1 Variation about mean shape, along first principal component*

Pedestrians are tracked by fitting this model to intensity discontinuities in successive frames, assuming constant linear motion to constrain the search within the framework of a Kalman filter. The results of tracking two pedestrians moving side-by-side are shown in Figure 2.

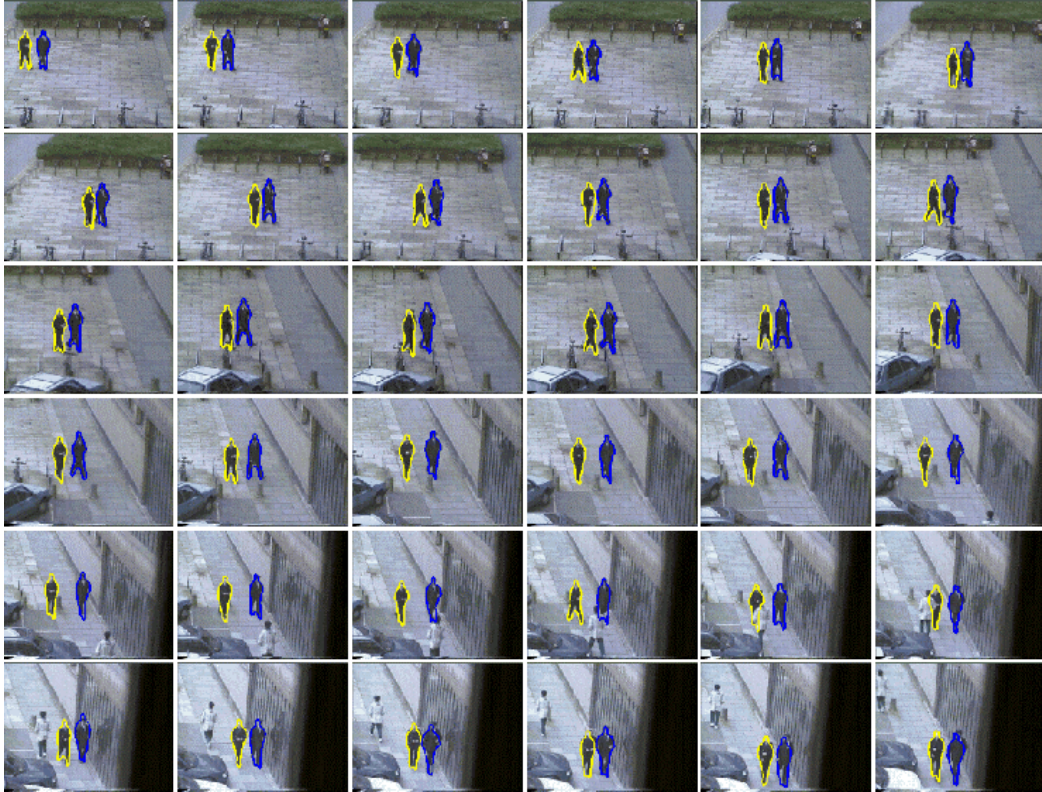


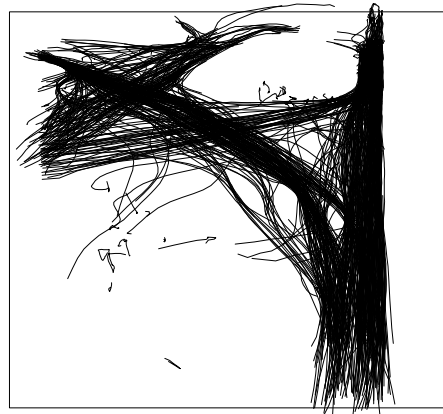
Figure 2 Typical output from the pedestrian tracker

### **Building the spatio-temporal model**

A set of 622 trajectories acquired from the scene shown in Figure 3(a) is depicted in Figure 3(b) - the lines show the path of the base of each pedestrian's profile in the image-plane.



(a)



(b)

Figure 3 (a) experimental site, (b) the set of trajectories extracted

The modelling of behaviour is in two stages. In the first stage, the local spatio-temporal behaviour of pedestrians is modelled in terms of their instantaneous locations and velocities within the image plane. This is achieved by discretising each trajectory into a sequence of locations with corresponding velocities - each represented by a 4-D vector  $(x, \dot{x}, y, \dot{y})$ , known as *state* vectors. The complete set of state vectors is represented by a small number of prototype states using vector quantisation. A set of 1000 prototype states derived from the trajectories shown in Figure 3(b) is shown in Figure 4.

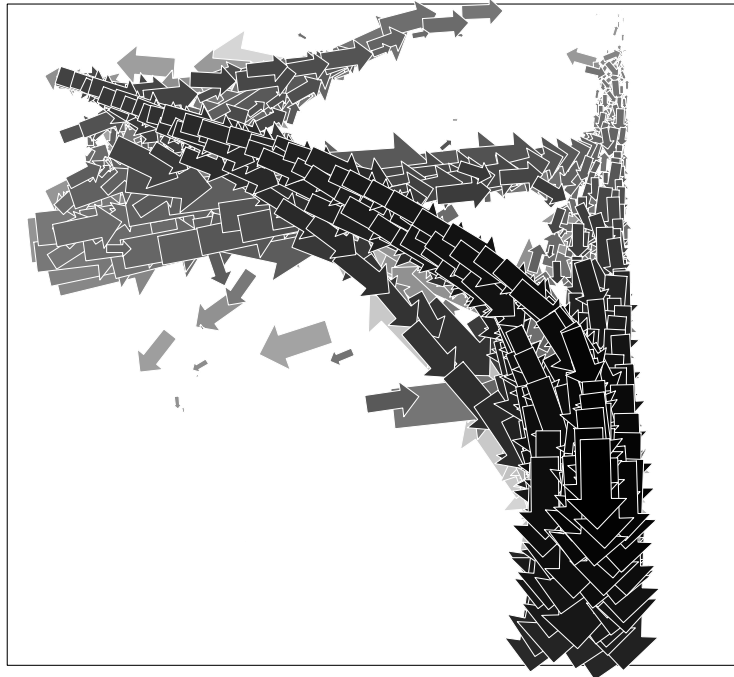
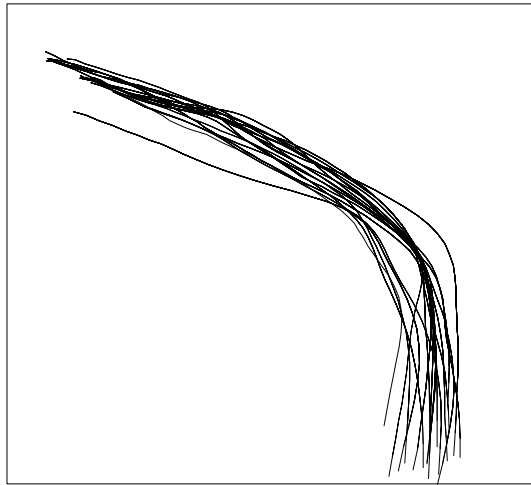


Figure 4 State prototypes representing location (position of arrow) and velocity (size and direction of arrow).

The set of state prototypes provide a simple form of model for the behaviour of pedestrians in a scene. For example, a limited form of extrapolation is possible, although the accuracy is limited by the short-time span of the temporal memory. What happens next is dependent only on the current state and not on any previous states.

The second stage builds on the first to provide a vector representation for extended behaviours. The way in which this is achieved may be visualised as follows. Each state prototype has an associated activation level that is initialised to zero for the encoding of a single trajectory. The trajectory is traced from beginning to end, and at each time step the activation of each state prototype either decays by a fixed proportion or takes on a value that is a linearly decreasing function of the prototype's distance from the current state, whichever gives the largest value. The pattern of activation levels at each time-step provides an encoding of the behaviour up until that point on the trajectory and this is recorded as a vector - referred to as a *behaviour vector* to distinguish from the state vectors.

The same procedure is repeated for all trajectories in the training set, giving a large collection of behaviour vectors encoding these trajectories and all partial trajectories implicit in their generation. This approach to encoding the evolution of a system is equivalent to the so-called leaky neural network model. The shading of state prototypes in Figure 4 shows their activation levels after a trajectory similar to those shown in Figure 5 has been completely traversed.



*Figure 5 Training trajectories nearest to a single behaviour prototype*

Finally, the distribution of behaviour vectors extracted from the training data are themselves encoded by a set of prototype behaviours, again derived by vector quantisation. For the data shown in Figure 3, 1000 behaviour prototypes were found to provide a sufficiently accurate encoding. Figure 5 shows all those trajectories from the training set that are closest to one of these prototypes.

The final behaviour prototypes provide a compressed model for the range of behaviours observed in the training video segment. This model can be adapted to serve as a piecewise uniform probability density function in which each prototype is replaced by a uniform region with magnitude proportional to the local density of prototypes, which is in turn proportional to the observed density of training behaviours (for details see Johnson and Hogg, 1996).

This probabilistic model can be used in several ways. The typicality of a given pedestrian trajectory may be determined directly from the corresponding density. Within a real-time system, the typicality may be evaluated frame by frame to provide a running assessment based on the current trajectory. Figure 6 shows a single frame from a video sequence in which typicality is high for two people moving along a well trodden path (indicated by superimposed circles) and low for someone moving in an unfamiliar way as they remove a cycle chained to railings (indicated by a superimposed triangle).



*Figure 6 Typicality assessment for three pedestrians*

Unfortunately, the behaviour model is not easily generative in the sense that it might be used to extrapolate forward partial trajectories or indeed to produce original, but characteristic, animations. Although the information required is contained in the representation, it is not easily extracted. To rectify this problem, a Markov chain is superimposed on top of the behaviour prototypes with transitions defining the ways in which behaviours in the neighbourhood of prototypes may evolve between frames - in general, extending the history of states by the addition of a new frame changes the associated behaviour. The probability of a transition is derived from the proportion of such transitions observed in the training set.

Although this extension to the behaviour model satisfies the Markov property that the probability of moving to the next prototype is dependent only on the current prototype, because the current prototype encodes the history within itself, there remains a longer duration temporal dependence. This is not simply a Markov chain over the 4-D state prototypes.

To extrapolate forwards a trajectory, the behaviour defined by the history up until the current frame is itself extrapolated forwards by taking the most probable transitions in the Markov chain, starting with the nearest prototype to the current behaviour. The state of a moving person (i.e. location and velocity) is recovered from each of these behaviour vectors by selecting the state prototype corresponding to the most active element. Finally, a smooth trajectory is interpolated between these state prototypes.

## ***Models of Interaction***

The previous section has described a method for modelling the behaviour of individuals. Our main aim in this paper is to show how this framework may be applied for the visual simulation of a participant in simple kinds of interaction involving two people. For example, we wish to simulate the facial expressions of a plausible partner in animated conversation. In our current work, we have focused on handshakes between two people viewed from the side. On being shown an individual with hand out-stretched, we wish to generate the image of a plausible partner to complete the hand-shake.

A novel solution is proposed (Johnson, Galata and Hogg, 1998) in which a probabilistic model of the joint behaviour of pairs of individuals is first acquired by observing sets of typical interactions. To simulate a virtual partner in an interaction, the model is partially instantiated by the appearance of a real partner and the virtual partner chosen to be the most likely completion of the interaction in a probabilistic sense.

As a first step to achieving this, a model for the target behaviour is constructed as before. A video segment containing many examples of the target behaviour is recorded to be processed off-line. Figure 7 shows a single frame from such a training set. Profiles of the two individuals in the training video are extracted by subtracting successive frames from a reference image of the background, obtained by median filtering a fixed number of previous frames on a pixel by pixel basis. Details of this process can be found in Baumberg and Hogg (1996). Thresholding the difference images partitions the frames into foreground regions corresponding to the profile of the two individuals and stationary background. For the collection of training data we have not found it necessary to use a more robust model-based tracker.

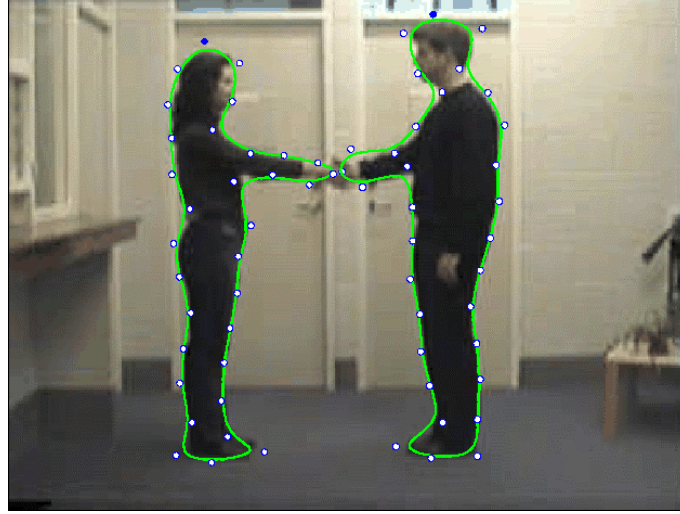


Figure 7 Single frame from training set of handshakes. The locations of B-spline control points are shown around each profile.

Once the profiles have been obtained, a 3<sup>rd</sup> order B-spline contour with a fixed number of control points is wrapped around each of these independently. In reality, the profiles will form a connected region which must be separated beforehand into two disjoint regions by simply dividing along a line between the two individuals - this is not a generally applicable procedure but one that serves our present purpose. Figure 7 shows the spline contour wrapped around a single frame, with control points marked. The contour is uniquely represented by the concatenation of these control points:

$$\mathbf{S} = (x_1, y_1, x_2, y_2, \dots, x_n, y_n)$$

The way in which a behaviour model is obtained now mirrors the way in which such a model was obtained for trajectories, but with the 4-D state vectors (location and velocity) replaced by the concatenation of the profiles for the left and right individuals ( $\mathbf{S}^L, \mathbf{S}^R$ ), their separation and their relative height - together with the first derivative of these parameters.

$$\mathbf{F} = (\mathbf{S}^L, \dot{\mathbf{S}}^L, \mathbf{S}^R, \dot{\mathbf{S}}^R, s, \dot{s}, h, \dot{h})$$

$s$  is separation

$h$  is relative height

For the handshake experiments, there are 32 control points for each profile giving a state vector of size 260 ( $2 \times 32 \times 2 \times 2 + 2 \times 2$ ). From the training data, 200 prototype states are derived by vector quantisation. From these, are generated 400 prototype behaviours linked into a Markov chain. This is our behaviour model for the act of handshaking.

As before, the behaviour model may be used in several ways. We can extrapolate forwards from a segment of tracked behaviour. This is illustrated in Figure 8 where the most likely sequence of future profiles (determined by the model) is superimposed on the final frame from an input sequence. These profiles provide a plausible future for the interaction, although of course the real interaction may have developed differently. In general, it is possible to generate different extrapolations probabilistically from the model.

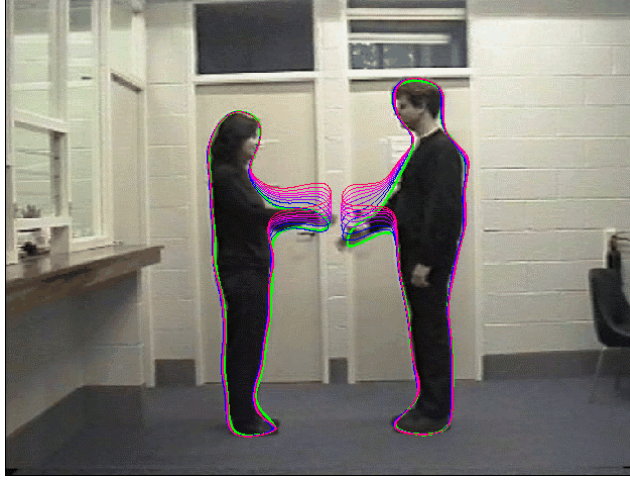


Figure 8 Predicting future behaviour

The model may also be used to extrapolate a behaviour that is partially occluded and to fill in the missing parts. In doing this, we must deal with uncertainty in the elements of the state vector acquired from pre-processing each incoming frame. A Bayesian framework is adopted in which the posterior density for the hypothesised state  $\hat{\mathbf{F}}_t$  at each frame is estimated recursively from a prior density for the state and the likelihood for the current state observation  $\mathbf{F}_t$ :

$$P(\hat{\mathbf{F}}_t | \mathbf{F}_t, \dots, \mathbf{F}_0) \propto P(\mathbf{F}_t | \hat{\mathbf{F}}_t) P(\hat{\mathbf{F}}_t | \mathbf{F}_{t-1}, \dots, \mathbf{F}_0)$$

The prior density is obtained by extrapolation from the posterior for the previous time step using the Markov model.

Our particular interest is in dealing with the complete occlusion of one of the participants in an interaction - without loss of generality we suppose this person stands on the right of the handshake. For this purpose, the likelihood function depends only on the Euclidean distance between the vectors representing the observed and hypothesised profiles of the left-hand person:

$$P(\mathbf{F}_t | \hat{\mathbf{F}}_t) = \exp\left(-\frac{\|\hat{\mathbf{S}}_t^L - \mathbf{S}_t^L\|^2}{2\sigma^2}\right)$$

The posterior density itself is represented by the distribution of a fixed number of multiple state hypotheses. This overall framework for representing and updating the posterior density is essentially the CONDENSATION algorithm (Blake and Isard, 1998).

In general, the maximum of the posterior density provides a plausible final hypothesis of the current state of an interaction. Unfortunately, it is hard to estimate this from the representation of the density. Instead the state hypothesis that maximises the likelihood function is selected. This turns out to satisfy our purposes in practice.

Several frames from a sequence in which an occluded partner is reconstructed in this way are shown in Figure 9. Since processing is currently off-line, the visible person shown in this figure is in fact shaking hands with a real person to maintain natural movements throughout the action. The sequence has been pre-processed to remove this person and thereby to simulate an occlusion.

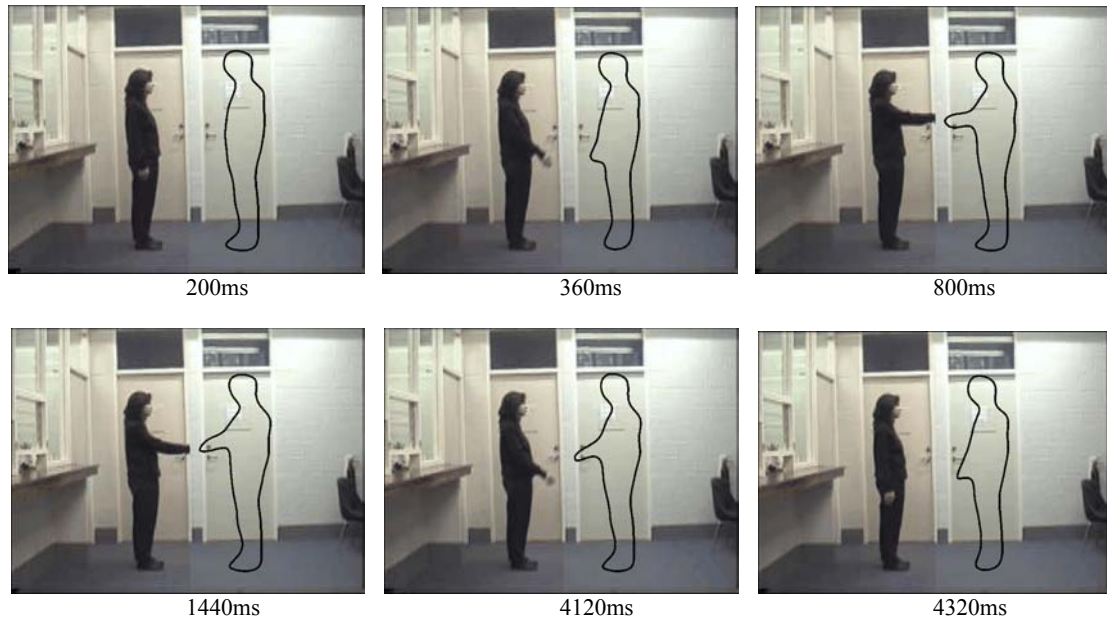


Figure 9 Interaction with a virtual person

The interest in this example is not only in dealing with occlusion but in the possibility for using the same approach to simulate a virtual partner in the interaction. To illustrate this idea in a different domain, we have carried out preliminary experiments with a virtual face, trained to respond to facial movements of a user with appropriate responses. A joint model of two faces interacting with one another was constructed from the output of a facial tracker. For these experiments, the visual face tracker described by Edwards, Taylor and Cootes (1998) has been implemented. Figure 10(a) shows a single frame from a training sequence in which the right-hand individual is nodding their head in response to head-shakes of the left-hand person. Figure 10(b) shows a virtual face on the right-hand side, nodding in response to head-shakes from the real face on the left-hand side.

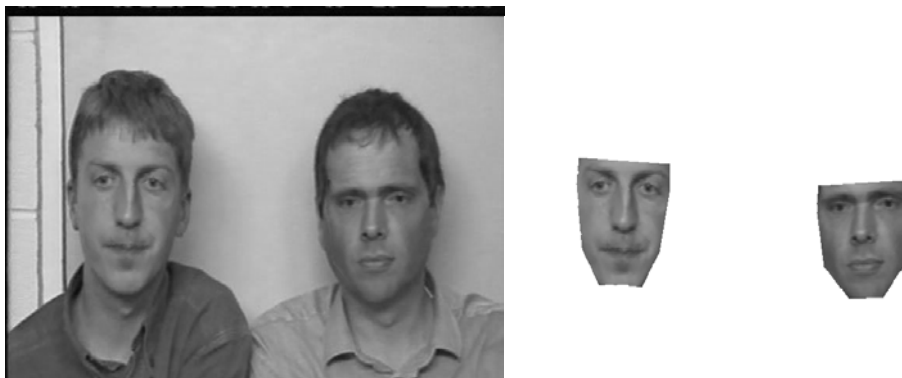


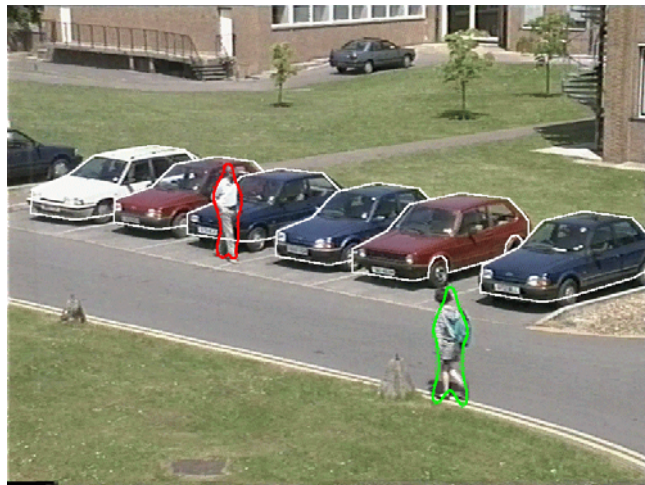
Figure 10 (a) Tracked faces for behavioural training data (b) reconstructed face (right) responding to a real face (left).

## Car Park Surveillance

This section deals with a different kind of application for the overall statistical approach to behaviour characterisation developed above - that of car park surveillance.

Detecting atypical behaviour in car parks depends on building some kind of understanding of the relationship between people in the car park and the cars themselves. A simple characterisation of locational behaviour is insufficient since this will depend crucially on the positions of vehicles in the car park. When the car park is empty, people are free to cross parking areas to reach the exit, whereas a full car park causes people to navigate around cars to reach their destination. Nevertheless, the relative location of people and vehicles seems to be important. Someone visiting several cars and remaining stationary at each may appear to be suspicious to the human eye. In line with earlier work, we might wish to detect such behaviours as outliers from a model for normal behaviour represented in terms of the relative motion between cars and people and learnt automatically from the observation of a target car park scene over several hours of activity. We go some way towards this ideal by introducing a statistically calibrated measure of interest in an individual as they traverse a car park, and employing a supervised learning procedure to distinguish between normal and unusual behaviours.

The prototype system is based on the model-based tracking procedure of Baumberg and Hogg (1994,1996) combined with a 3-D model-based tracking system for vehicles (Tan, Sullivan and Baker, 1994). Vehicles are tracked first, and occlusion masks indicating the depth and extent of projected vehicles within the image plane are passed to the person tracker to improve its reliability (Remagnino and Baumberg et al. 1997). This greatly improves tracking performance within car parks, since people are often partially occluded by vehicles. The image-plane position of the base of each tracked person is back-projected onto the ground plane to provide the final output of a series of trajectories of people and vehicles in 3-D. A result of the combined tracker is shown in Figure 11.



*Figure 11 Combined tracker for people and vehicles*

In one experiment at a University car park, we recorded two hours of video at the start of the working day, during which time 129 people were observed to enter and leave the scene, of which 11 were behaving unusually and performed by an actor. Figure 12 shows a view of this car park during one of the unusual scenarios - the path followed by the actor is shown.



Figure 12 Single frame depicting an unusual behaviour

Figure 13 shows a schematic view of the same car park seen from above. Vehicles are depicted by rectangles and the path of the pedestrian by a curved line. In the top row, pedestrians are moving along a pathway running alongside the car park. The trajectories shown in the second and third rows are typical of car park behaviour, including several examples of people leaving vehicles and moving directly to the exit. The bottom row shows five of the unusual behaviours recorded.

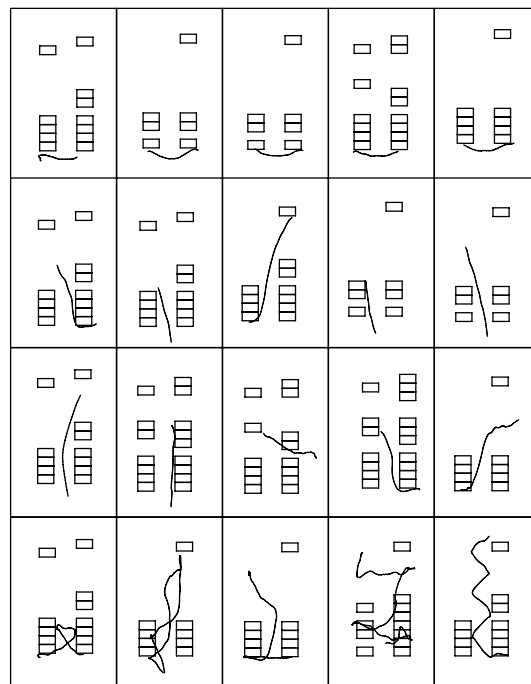


Figure 13 Schematic aerial view showing 20 paths recorded from experimental video. The cars are depicted by rectangles and the path followed by a curved line.

An important aspect of the behaviour of someone in a car park is their relationship with the vehicle to which they are closest at any time. The time during which they are in the car park can be partitioned into separate intervals during which they are closest to a different vehicle. After trying several different representations for behaviour, the most useful characterisation for our purposes was the distribution of two measurements taken at the point at which a person is at the closest point of approach during each interval.

The two measurements recorded are the current distance separating the person from the vehicle and the speed of the person. When someone revisits a vehicle after being closer to another vehicle in the intervening period, measurements will be recorded for both intervals separately. The graph in Figure 14 shows the distance of a single pedestrian to the nearest car during a period of around 40 seconds containing three complete intervals. The sharp peaks mark the transitions between intervals and measurements are recorded at the minimum of the curve between each peak.

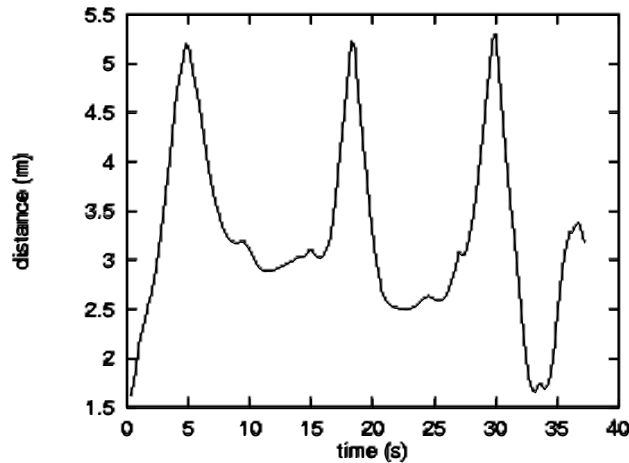


Figure 14 The variation in distance between a pedestrian and the nearest car

Figure 15 shows the scatter plot of speed-distance measurements acquired from 129 trajectories. These data are summarised as a cumulative probability distribution represented as a 2-D array of values, shown in Figure 16.

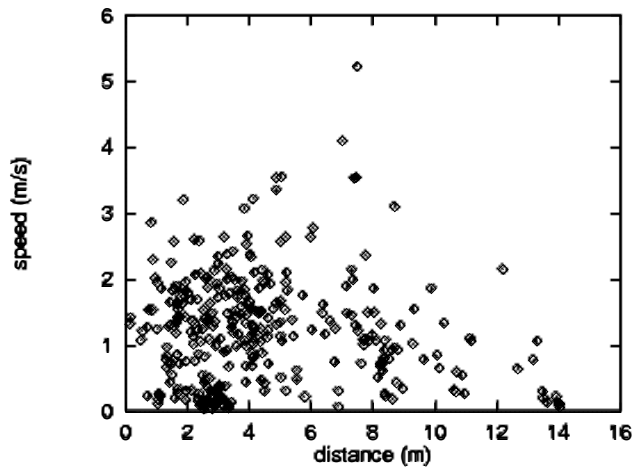


Figure 15 Scatter plot of speed-distance measurements at closest approach

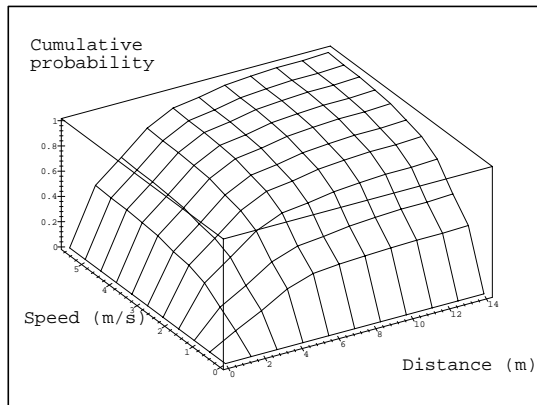


Figure 16 Cumulative probabilities for speed and distance at closest point

A measure of significance is obtained directly from a single point of closest approach by looking up the associated cumulative probability. The closer a person is to the vehicle and the slower they move, the more interesting the event. In order to distinguish between normal and unusual behaviour, the cumulative probabilities of all points of closest approach for an individual are ordered, and the smallest five values retained  $\{p_i | i = 1,5\}$ . A planar decision surface is used to distinguish between normal and unusual behaviour:

$$\text{IF } \sum_{i=1}^5 a_i p_i > 0.5 \text{ THEN unusual ELSE normal}$$

The sequences of smallest probabilities are shown in Figure 17 for two normal behaviours and in Figure 18 for two unusual behaviours.

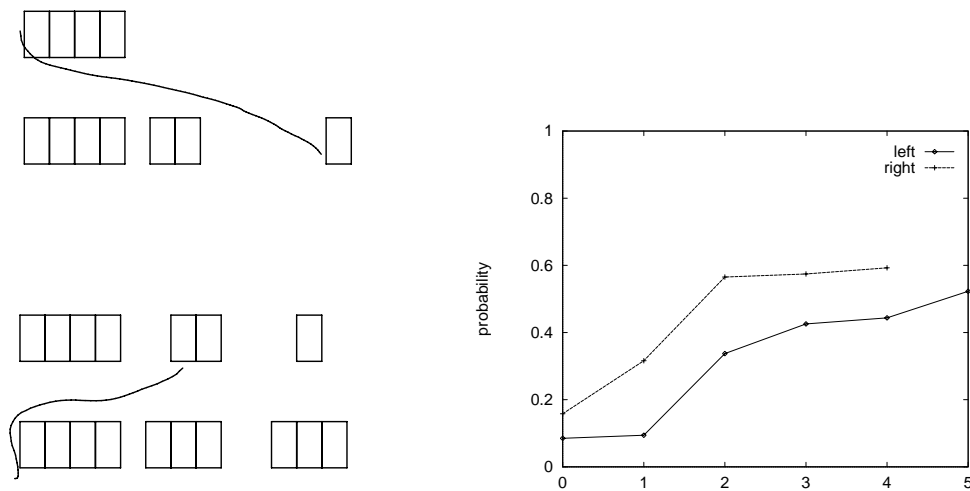


Figure 17 Two normal views and their sequence of smallest probabilities

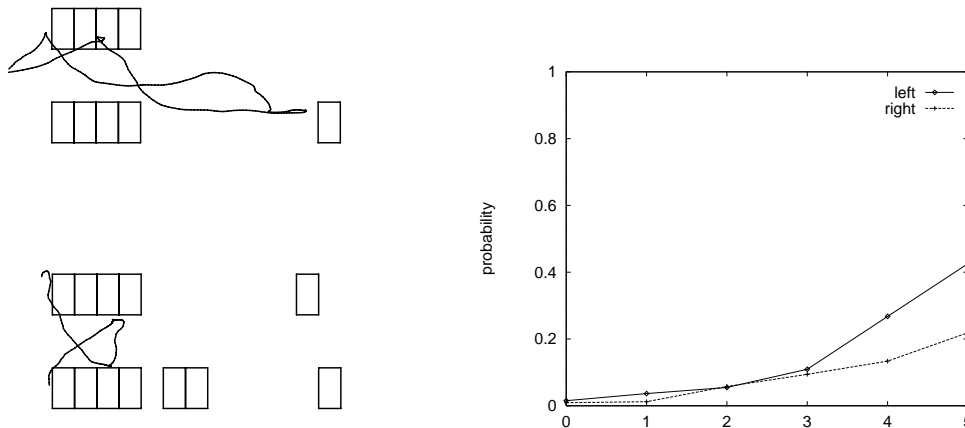


Figure 18 Two unusual views and their sequence of smallest probabilities

In an experiment, the decision surface (i.e. the values of  $a_i$ ) was selected to partition a training set of 59 observed trajectories of which 5 were unusual. A test set of 70 trajectories, 6 of which were unusual, was classified using the same decision surface. All 6 unusual behaviours and all but 4 of the normal behaviours were classified correctly.

## References

- Baumberg, A. and Hogg, D.C. (1994), An Efficient Method for Contour Tracking using Active Shape Models, in Proc. of IEEE Workshop on Motion of Non-Rigid and Articulate Objects (Austin, Texas).
- Baumberg, A.M. and Hogg, D.C. (1996), Learning Deformable Models for Tracking the Human Body, in Motion-Based Recognition, M. Shah and R. Jain. (eds.), Kluwer Academic, 39-60.
- Blake, A. and Isard, M. (1998), Active Contours, Springer-Verlag.
- Cootes, T.F., Cooper, D., Taylor, C.J. and Graham, J. (1995), Active Shape Models - Their Training and Application. Computer Vision and Image Understanding. 61(1), 38-59.
- Edwards, G.J., Taylor, C.J., and Cootes, T.F. (1998), Interpreting Face Images using Active Appearance Models, Face and Gesture Conf.
- Johnson, N. and Hogg, D.C. (1996), Learning the distribution of object trajectories for event recognition, Image and Vision Computing 14(8), 609-615.
- Johnson, N., Galata, A. and Hogg, D.C. (1998), The Acquisition and Use of Interaction Behaviour Models, Proceedings of IEEE Conf. On Computer Vision and Pattern Recognition, 866-871.
- Morris, R.J. and Hogg, D.C. (1998), Statistical Models of Object Interaction, Proc. IEEE Workshop on Visual Surveillance, Bombay, 81-85 (to appear in the International Journal of Computer Vision).
- Remagnino, P., Baumberg, A., Grove, T., Hogg, D., Tan, T., Worrall, A. and Baker, K. (1997), An integrated traffic and pedestrian model-based vision system, in Proc. British Machine Vision Conference 1997, 380-389.
- Tan, T.N., Sullivan, G.D. and Baker, K.D. (1994), Recognising objects on the ground-plane. Image and Vision Computing 12, 164-172.
- Terzopoulos, D. and Waters, K. (1990), Physically-based facial modelling, analysis, and animation. The Journal of Visualization and Computer Animation, 1(2), 73-80.